Sociology 7706: Longitudinal Data Analysis
Instructor: Natasha Sarkisian

Introduction to Longitudinal Data Analysis

Types of data:

- Cross-sectional data = data collected at one point in a time
- Longitudinal data = data collected on the same variables from the same units are measured at 2 or more time periods
- Quasi-longitudinal data:
  - Repeated cross-sections: same variables are measured at 2 or more periods, but on different units
  - Time-ordered cross-sections: data are collected from the same units at two or more periods, but each variable is measured only once
  - Retrospective studies: data about different time points in the past are collected all at once

Advantages of longitudinal data:

- Can examine patterns of change over time
- Can study individual development/trajectories
- Can analyze duration
- Can help locate the causes of social phenomena

Basic conditions for establishing causality:

1. Correlation (X and Y must change together)
2. Theory (logical explanation)
3. Non-spuriousness (other explanations must be ruled out)
4. Temporal order (X must precede Y temporally)

Longitudinal research can help ensure temporal order, but by itself does not assure causality.

Disadvantages of longitudinal data:

- Costly and time-consuming
- Panel attrition problems: Refusals, changes of residence, death
- Discrete time measurement: Exactly what happens between the time points is unknown
- Time lag problems: Time intervals might not match the lag between cause and its consequence
- Panel conditioning: Responses in one wave can be influenced by those given in the previous waves, respondents themselves can change as a result of participating in the study

Longitudinal analysis:

Observations on the same unit over time → not independent → have to apply special techniques

Two types of questions:

- Descriptive: What kind of change takes place?
- Explanatory: What predicts this change (its nature and timing)?

Notation:

Y_{it}

$i = 1 \ldots N$ -- units

$t = 1 \ldots T$ – time points

Example: T=2, N=3

ID	Time	X	Y
1	1	x_{11}	y_{11}
1	2	x_{12}	y_{12}
2	1	x_{21}	y_{21}
2	2	x_{22}	y_{22}
3	1	x_{31}	y_{31}
3	2	x_{32}	y_{32}

This is long format – one row is called "person-year" or "country-year", etc.

Same dataset in wide format:

ID	X1	Y1	X2	Y2
1	x_{11}	y_{11}	x_{12}	y_{12}
2	x_{21}	y_{21}	x_{22}	y_{22}
3	x_{31}	y_{31}	x_{32}	y_{32}

- N small, T large → Cross-sectional time series (common when units are countries)
- N large, T small → Panel data (common when units are individuals)

Panel data can be balanced when all individual cases are observed in all time periods or unbalanced when individual cases are not observed in all time periods.

Analytical Models and Data Structures

FEW CASES *(n < 20)* *MANY PERIODS* *(t > 20)*	*MANY CASES* *(n > 100)* *MANY PERIODS* *(t > 10)*
ARIMA models: covariates, transfer function models, interrupted time series models	Continuous time event history analysis: Cox proportional hazards and parametric hazard models
Autoregressive (AR) time series models	
Lagged endogenous variable (LEV) models	Multilevel growth curve models
Multivariate dynamic analysis of categorical data with optimal scaling	
FEW CASES *(n < 20)* *FEW PERIODS* *(t < 10)*	*MANY CASES* *(n > 100)* *FEW PERIODS* *(t < 10)*
Pooled cross-sectional/time-series analysis	Linear panel analysis conditional change model (lagged endogenous variable)
	Linear panel analysis unconditional change model (change score)
	Latent growth curve analysis
	Discrete time event history analysis
	Multilevel growth curve models

From: "Longitudinal Research" by Scott W. Menard